Understanding the role of individual units in a deep neural network

Year: 2020 | Citations: 494 | Authors: David Bau, Jun-Yan Zhu, Hendrik Strobelt, Àgata Lapedriza, Bolei Zhou

Abstract

Deep neural networks excel at finding hierarchical representations that solve complex tasks over large datasets. How can we humans understand these learned representations? In this work, we present network dissection, an analytic framework to systematically identify the semantics of individual hidden units within image classification and image generation networks. First, we analyze a convolutional neural network (CNN) trained on scene classification and discover units that match a diverse set of object concepts. We find evidence that the network has learned many object classes that play crucial roles in classifying scene classes. Second, we use a similar analytic method to analyze a generative adversarial network (GAN) model trained to generate scenes. By analyzing changes made when small sets of units are activated or deactivated, we find that objects can be added and removed from the output scenes while adapting to the context. Finally, we apply our analytic framework to understanding adversarial attacks and to semantic image editing.